Home Work 4

Harshith Bondada
M10724854

1. Read PGA data into R (PGA.csv). Below is the description of variables.

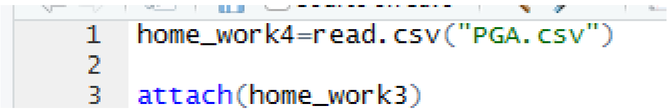
Source: sportsillustrated.cnn.com

Description: Performance statistics and winnings for 196 PGA participants during, 2004 season.

Variable: Name, Age, Average Drive (Yards), Driving accuracy (percent), Greens on regulation (%), Average # of putts, Save Percent, Money Rank, # Events, Total Winnings ($), Average winnings ($).

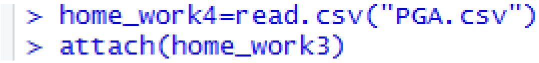
We read the data using the read.csv function.

We use the variable home_work 4 to store the data.



```
home_work4=read.csv("PGA.csv")

attach(home_work3)
```

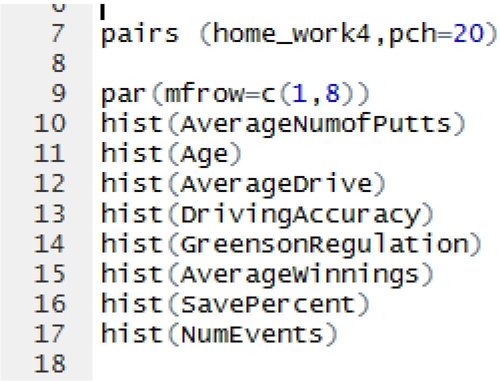
This shows that the operation is successful.



```
> home_work4=read.csv("PGA.csv")
> attach(home_work3)
```

2. Visualize the data using scatter plot and histogram.

We use the function pairs to plot the scatter plot and hist function to draw the histogram.

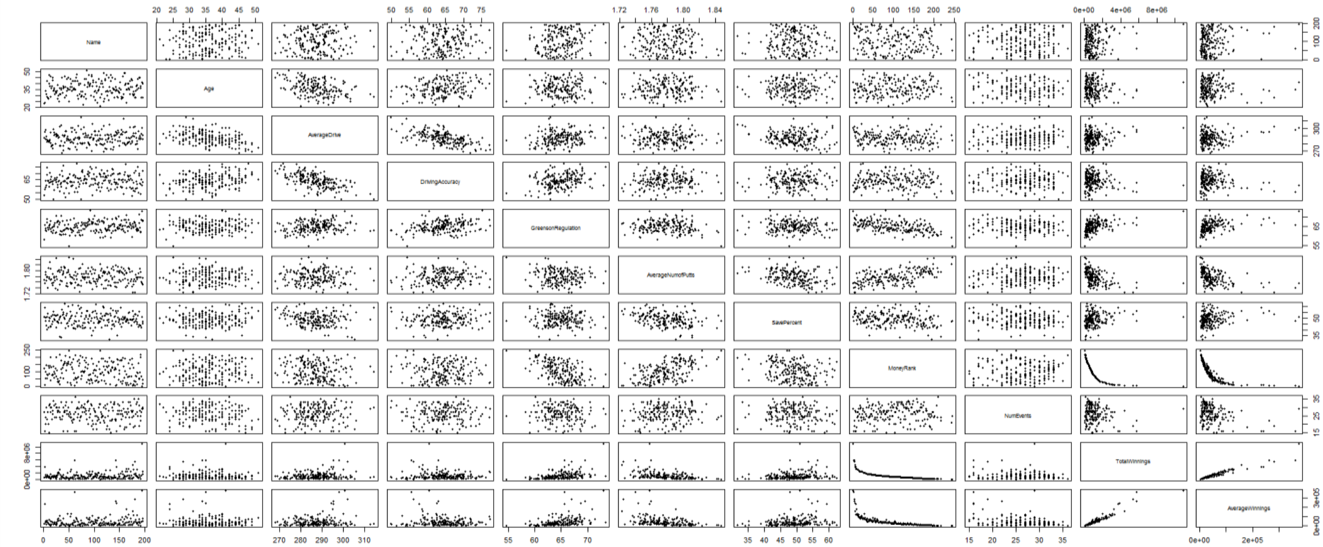


```
pairs (home_work4,pch=20)

par(mfrow=c(1,8))
hist(AverageNumofPutts)
hist(Age)
hist(AverageDrive)
hist(DrivingAccuracy)
hist(GreensonRegulation)
hist(AverageWinnings)
hist(SavePercent)
hist(NumEvents)

```

Home Work 4

Harshith Bondada
M10724854

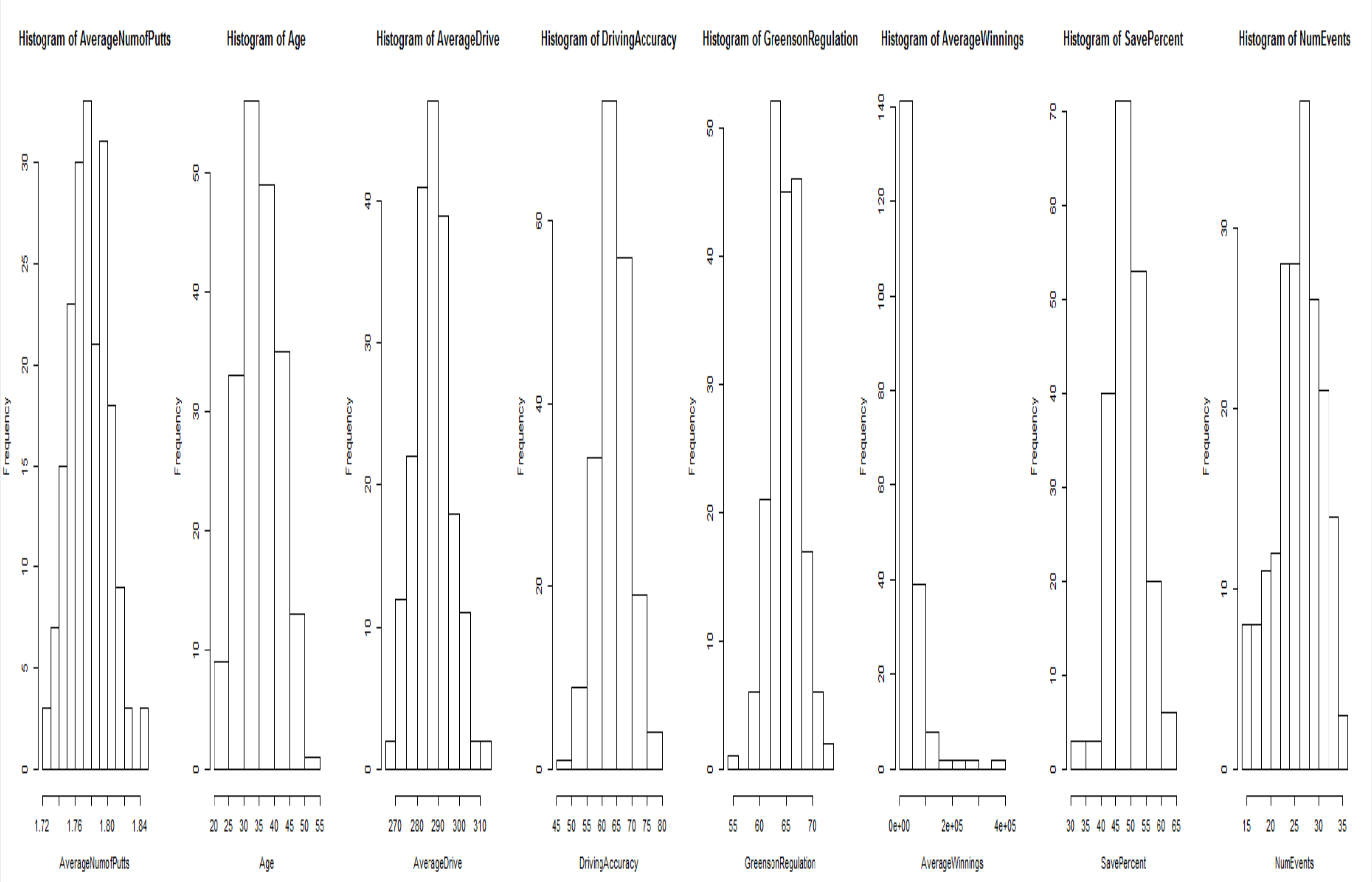
This is the scatter plot for the data in the data set.



Home Work 4

Harshith Bondada
M10724854

These diagrams are the histograms for the datset.



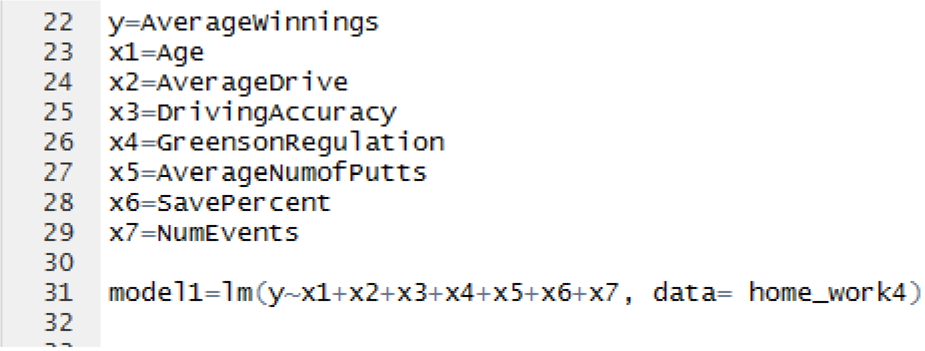
Home Work 4

Harshith Bondada
M10724854

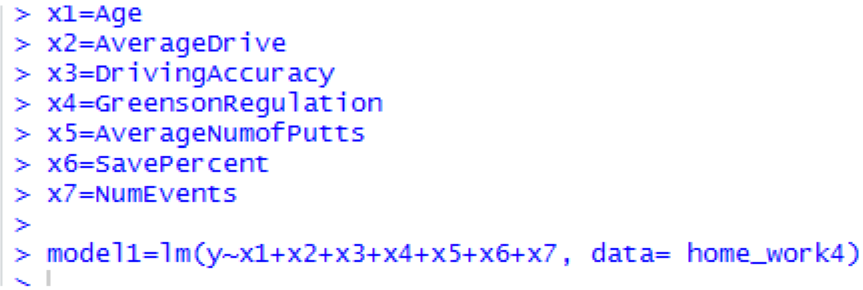
3. Build a linear regression using Average winnings as response variable and using Age, Average Drive (Yards), Driving accuracy (percent), Greens on regulation (%), Average # of putts, Save Percent, and # Events as covariates.

We store the response variable in y and the remaining co variates in x_j. j means 1 to 7.

We generate a model1 for the regression.

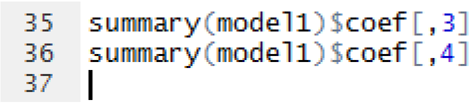


```
y=AverageWinnings
x1=Age
x2=AverageDrive
x3=DrivingAccuracy
x4=GreensonRegulation
x5=AverageNumofPutts
x6=SavePercent
x7=NumEvents

model1=lm(y~x1+x2+x3+x4+x5+x6+x7, data= home_work4)

```



```
> x1=Age
> x2=AverageDrive
> x3=DrivingAccuracy
> x4=GreensonRegulation
> x5=AverageNumofPutts
> x6=SavePercent
> x7=NumEvents
>
> model1=lm(y~x1+x2+x3+x4+x5+x6+x7, data= home_work4)
>
```

4. Perform t tests for these coefficient estimates. Obtain t statistics and p values, interpret the results, make a conclusion (i.e. reject or not reject) and explain why. Note: please explain what the null hypothesis is.

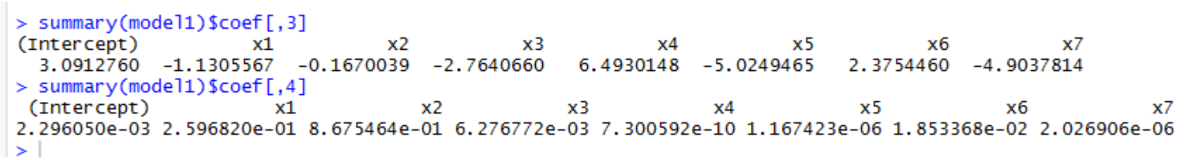


```
summary(model1)$coef[,3]
summary(model1)$coef[,4]

```

We do not reject the null hypothesis for Age and Average Drive as they are not significant, and their slope is not significantly different from zero. Whereas, for the other regressors we reject the null hypothesis as they are significant in explaining the variation in response variable, their slope coefficient is significantly different from zero.

Home Work 4

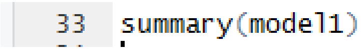
Harshith Bondada
M10724854



```
> summary(model1)$coef[,3]
(Intercept)          x1          x2          x3          x4          x5          x6          x7 
  3.0912760  -1.1305567  -0.1670039  -2.7640660   6.4930148  -5.0249465   2.3754460  -4.9037814 
> summary(model1)$coef[,4]
 (Intercept)           x1           x2           x3           x4           x5           x6           x7 
2.296050e-03 2.596820e-01 8.675464e-01 6.276772e-03 7.300592e-10 1.167423e-06 1.853368e-02 2.026906e-06 
> |
```

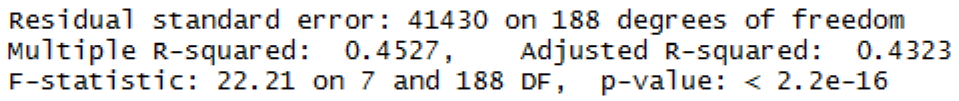
5. Use F test to test the significance of the regression. Obtain the F statistic and p value, interpret the results and make a conclusion.

We can get the F test by using the summary function. F value and p value are found at the bottom of the summary output.



```
summary(model1)
```

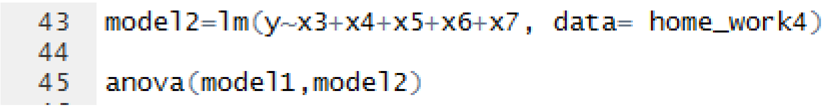
We find the p value to be less than 5% i.e., 0.05. Hence, we reject the null hypothesis that all slopes are zero.



```
Residual standard error: 41430 on 188 degrees of freedom
Multiple R-squared:  0.4527,	Adjusted R-squared:  0.4323 
F-statistic: 22.21 on 7 and 188 DF,  p-value: < 2.2e-16
```

6. Use a partial F test to test for two variables Age and Average Drive (Yards) together. According to your results, what do you conclude? Similarly, use the partial F test to test for three variables Age, Average Drive (Yards), and Save Percent together, what do you conclude?

We use the model 2 for testing the Age and Average Drive regressors.

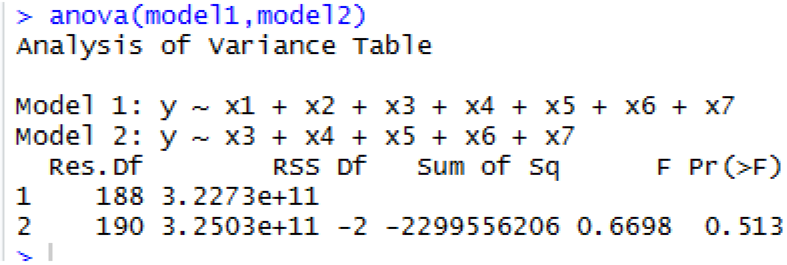


```
model2=lm(y~x3+x4+x5+x6+x7, data= home_work4)

anova(model1,model2)
```

We do not reject the null hypothesis. We cannot say that model 2 coefficients are significantly equal to zero. Therefore, we cannot remove all of those in favor of the simpler model.

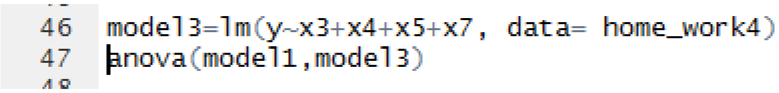
Home Work 4

Harshith Bondada
M10724854



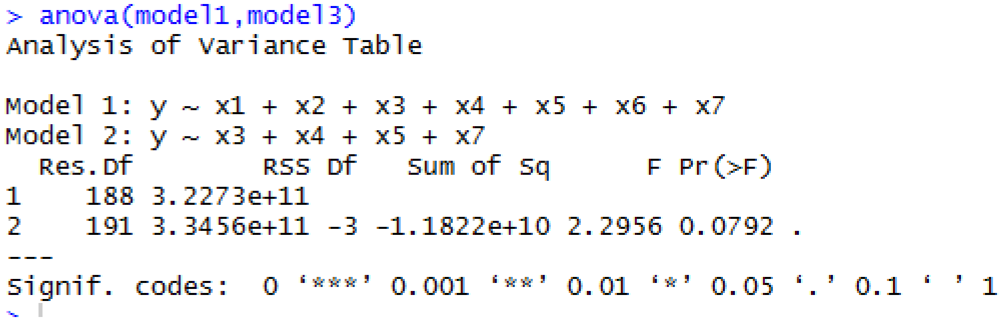
```
> anova(model1,model2)
Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7
Model 2: y ~ x3 + x4 + x5 + x6 + x7
  Res.Df        RSS Df   Sum of Sq      F Pr(>F)
1    188 3.2273e+11
2    190 3.2503e+11 -2 -2299556206 0.6698  0.513
> 
```

We use model 3 for testing Age, Average Drive, Save Percent regressors.



```
model3=lm(y~x3+x4+x5+x7, data= home_work4)
anova(model1,model3)
```

We do not reject the null hypothesis. We cannot say that model 2 coefficients are significantly equal to zero. Therefore, we cannot remove all of those in favor of the simpler model.



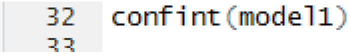
```
> anova(model1,model3)
Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7
Model 2: y ~ x3 + x4 + x5 + x7
  Res.Df        RSS Df   Sum of Sq      F Pr(>F)  
1    188 3.2273e+11                               
2    191 3.3456e+11 -3 -1.1822e+10 2.2956 0.0792 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

Home Work 4

Harshith Bondada
M10724854

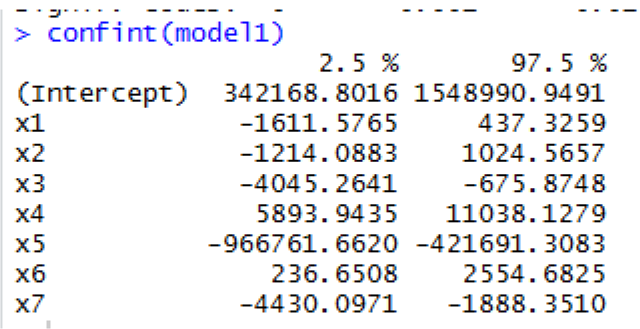
7. Obtain the interval estimation for all the intercept and slope coefficients.

We obtain the interval estimates using the confint function for all the intercept and slope functions.



```
confint(model1)
```

We get the output between the 95% probability range for the intercept and the slopes.



```
> confint(model1)
                    2.5 %       97.5 %
(Intercept)  342168.8016 1548990.9491
x1            -1611.5765     437.3259
x2            -1214.0883    1024.5657
x3            -4045.2641    -675.8748
x4             5893.9435   11038.1279
x5          -966761.6620 -421691.3083
x6              236.6508    2554.6825
x7            -4430.0971   -1888.3510
```

8. Using the regression in question 3, make a prediction for the case of:

Age = 35,

Average Drive = 287,

Driving Accuracy = 64,

Greenson Regulation = 64.9,

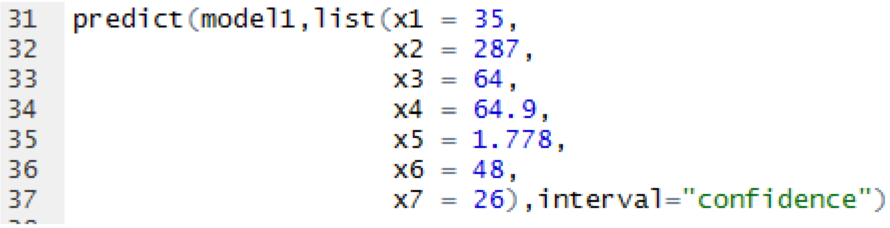
Average Num of Putts = 1.778,

Save Percent = 48,

Num Events = 26,

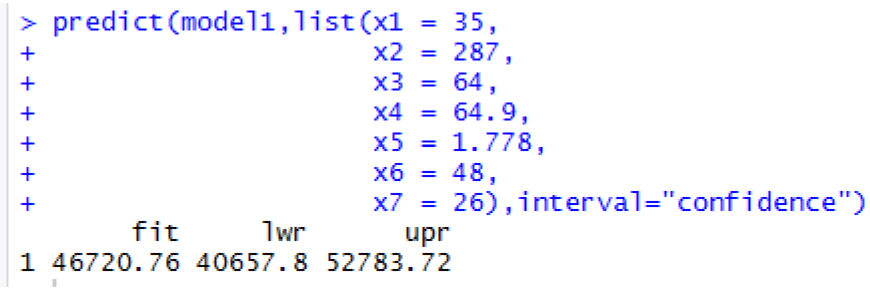
The prediction should include fitted value and interval estimation.

We use the predict function to get the prediction for new case**.**



```
predict(model1,list(x1 = 35,
                    x2 = 287,
                    x3 = 64,
                    x4 = 64.9,
                    x5 = 1.778,
                    x6 = 48,
                    x7 = 26),interval="confidence")
```

We can see that the predicted y^ is 46720.76 between the interval range of 40657.8 to 52783.72



```
> predict(model1,list(x1 = 35,
+                     x2 = 287,
+                     x3 = 64,
+                     x4 = 64.9,
+                     x5 = 1.778,
+                     x6 = 48,
+                     x7 = 26),interval="confidence")
       fit     lwr      upr
1 46720.76 40657.8 52783.72
```

9. Similarly, make another prediction for the case of

Age = 42,

Average Drive = 295,

Driving Accuracy = 69,

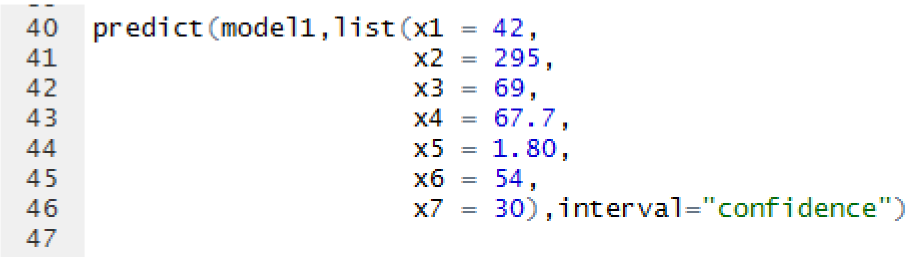
Greenson Regulation = 67.7,

Average Numof Putts = 1.80,

Save Percent = 54,

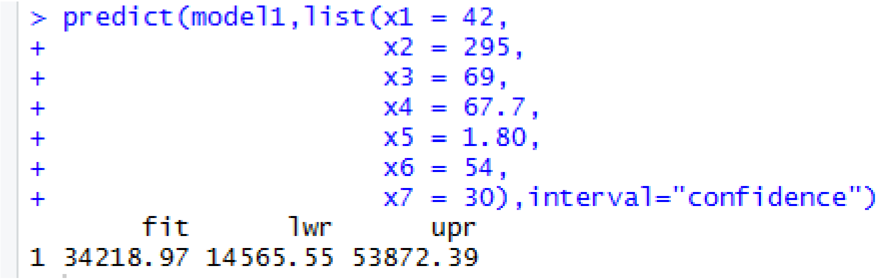
Num Events = 30,

The prediction should again include the fitted value and interval estimation. Compare the interval from question 8, what do you observe? For example, which interval is wider? And why?



```
predict(model1,list(x1 = 42,
                    x2 = 295,
                    x3 = 69,
                    x4 = 67.7,
                    x5 = 1.80,
                    x6 = 54,
                    x7 = 30),interval="confidence")
```

The interval in the question 9th is wider.



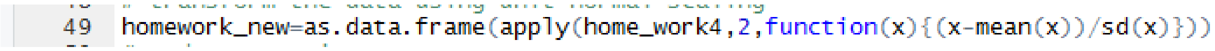
```
> predict(model1,list(x1 = 42,
+                     x2 = 295,
+                     x3 = 69,
+                     x4 = 67.7,
+                     x5 = 1.80,
+                     x6 = 54,
+                     x7 = 30),interval="confidence")
       fit      lwr      upr
1 34218.97 14565.55 53872.39
```

Home Work 4

Harshith Bondada
M10724854

10. Obtain the standardized regression coefficients and compare the influence of all variables.

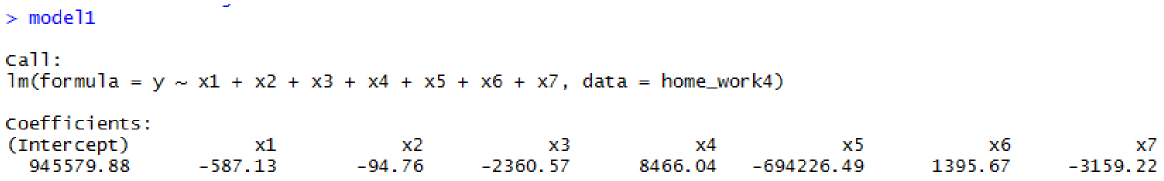
We first transform the data using the unit normal scaling



```
homework_new=as.data.frame(apply(home_work4,2,function(x){(x-mean(x))/sd(x)}))
```

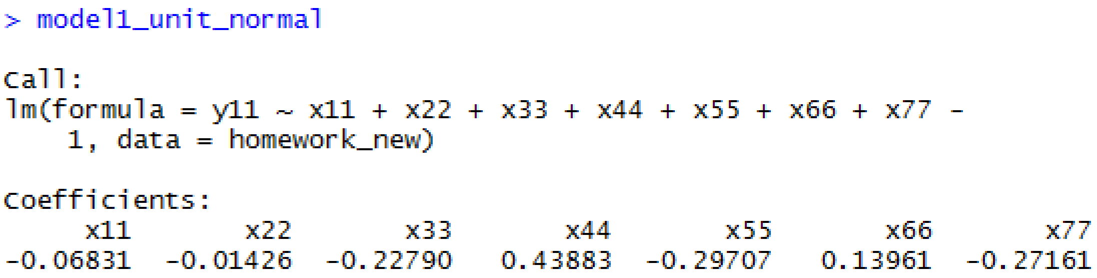
We first check the coefficients of model1 before scaling.

We find that the x4 has greater influence when compared to others.



```
> model1

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data = home_work4)

Coefficients:
(Intercept)           x1           x2           x3           x4           x5           x6           x7  
  945579.88      -587.13       -94.76     -2360.57      8466.04   -694226.49      1395.67     -3159.22  
```

After scaling we see that x44 which is same as x4 has greater influence after scaling.



```
> model1_unit_normal

Call:
lm(formula = y11 ~ x11 + x22 + x33 + x44 + x55 + x66 + x77 - 
    1, data = homework_new)

Coefficients:
     x11       x22       x33       x44       x55       x66       x77  
-0.06831  -0.01426  -0.22790   0.43883  -0.29707   0.13961  -0.27161  
```

This means that if the SD of x4 has increased by 1 SD of x4 then response variable should show a change of 0.43883 of that x4.